UNIVERSITI TEKNOLOGI MARA

MATHEMATICS AND STATISTICS

UNDERGRADUATE RESEARCH PROCEEDINGS 2025

UiTM CAWANGAN NEGERI SEMBILAN

AIR QUALITY ASSESSMENT IN KLANG USING PRINCIPAL COMPONENT ANALYSIS

Nur Afrina Syamimi Norazman[1], **Isnewati Ab Malek**[1,*]
[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Cawangan Negeri Sembilan, Kampus Seremban, 70300 Negeri Sembilan*
[*]isnewati@uitm.edu.my

Abstract

Air pollution, a major environmental concern, negatively affects human health and the environment. In Malaysia, urban areas such as Klang in Selangor have suffered from increased air pollution due to heavy industrial activities, dense populations, and high vehicular traffic, which are particularly vulnerable. This study identified Selangor's air quality trends using the Department of Environment (DOE) data. The monitoring station in Klang was selected based on six air pollutants (sulphur dioxide (SO_2), particulate matter below 10 microns (PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO) and particulate matter below 2.5 microns ($PM_{2.5}$) during three years from January 1, 2020, to December 31, 2022. This study aims to use the principal component analysis (PCA) method to classify the variables that predict the air pollution index (API) in Klang, Selangor. PCA was used to reduce dimensionality and identify significant pollutant predictors, with Bartlett's test and KMO measure supporting data suitability and rotating PCA reducing predictor variables. As a result, two principal components that highlighted PM_{10}, $PM_{2.5}$ and O_3 as the most important pollutants in Klang were revealed after the PCA was rotated using varimax rotation. The finding could assist DOE management in identifying the types of pollutants responsible for air pollution.

Keywords: Air Pollution Index, Principal Component Analysis (PCA), Klang

Introduction

The issue of air pollution is not new to any country worldwide, including Malaysia which is one of the most serious environmental issues facing the world today. There are plenty of different types of pollution, including noise, ground, water, and air pollution. Air pollution refers to the contamination of the atmosphere by another harmful substance. Air pollution is the contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere [1]. After considering the information regarding air pollution, it is not only dangerous for human health but also makes it unsuitable for people to live in if air pollution continues to increase in the future.

Motor vehicles, power plants, and industry as major sources of air pollution in Malaysia [2]. Industrial smoke, cigarette smoke, open burning, and household combustion devices contribute to air pollution. Moreover, [3] said that urban areas in Malaysia have the highest levels of PM_{10}, a harmful inhalable particle, and higher levels of (O_3), a secondary air pollutant produced by industrial emissions and motor vehicle exhausts. Malaysia has also seen an increase in carbon monoxide (CO) emissions due to fossil fuel reliance.

The US Environmental Protection Agency's Air Quality measure (AQI) is the most extensively used global air quality measure, adopted by many countries, including Malaysia [4]. Based on research from [5] said that the United States Environmental Protection Agency (US EPA) provides daily air quality conditions using six criteria: sulphur dioxide (SO_2), particulate matter below 10 microns (PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO) and particulate matter below 2.5 microns ($PM_{2.5}$). The index value set is classified as good, moderate, unhealthy, very unhealthy, hazardous, or emergency, indicating the air quality's potential harm to humans and the environment, as indicated in Table 1.

Table 1: Air Pollutant Index (API) of Malaysia

Air Pollution Index	Air Pollution Level
0 to 50	Good
51 to 100	Moderate
101 to 200	Unhealthy
201 to 300	Very Unhealthy
> 300	Hazardous
>500	Emergency

Seven million premature deaths annually are caused by air pollution. Malaysia's Ministry of Health reports that respiratory system diseases cause 10.36% of deaths, with 19.48% due to these diseases [1]. Pollutants contributing to air pollution include sulphur dioxide (SO_2), particulate matter below 10 microns (PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO) and particulate matter below 2.5 microns ($PM_{2.5}$). Each pollutant has unique effects on human health, with sulphur dioxide SO_2 causing asthma, particulate matter below 10 microns PM_{10} causing lung cancer, ozone O_3 causing breathing problems, nitrogen dioxide NO_2 related to asthma and other respiratory diseases, and carbon monoxide (CO) causing breathing difficulties, exhaustion, dizziness, and flu-like symptoms. Particulate matter below 2.5 microns $PM_{2.5}$ can also cause cardiovascular, cerebrovascular, and respiratory effects.

Accurate prediction of API is essential for implementing timely interventions, informed decision-making, and effective air quality management. However, identifying a robust model that balances accuracy, computational efficiency, and interpretability remains a challenge. This study aims to bridge this gap by leveraging Principal Component Analysis (PCA) for dimensionality reduction. By applying these advanced techniques, the research seeks to develop an optimized model for predicting API in Klang contributing to better air quality management and public health outcomes.

Methodology

Data Description

The study used secondary data from the Department of Environment's Malaysia air pollution index, specifically the Selangor index, covering the period from January 1, 2020 to December 31, 2022. The index was calculated using daily observations from the Klang air quality monitoring stations. Pollutants included sulphur dioxide (SO_2), particulate matter below 10 microns (PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO) and particulate matter below 2.5 microns ($PM_{2.5}$).

Data Processing

The data collection of the model for prediction in the current research was constructed using 1096 data received from six variables recorded daily from January 1, 2020 to December 31, 2022 from the Klang. The data were collected by the Department of Environment (DOE). Before proceeding to generate PCA, any absence of a dataset must be addressed with multiple imputations since PCA cannot handle the missing values directly. This method estimated a set of appropriate values for missing data based on the distribution of observed data.

Dimensionality Reduction

Principal component analysis (PCA) was a multivariate statistical method that combined information gathered from many different variables observed on the same individuals into a smaller number of variables known as principal components (PCs) [6]. According to [7], principal component analysis (PCA) was a dimensionality reduction (DR) technique that commonly utilised to reduce a large number of variables into a smaller number that still maintained some of the details in the large set. The PCA was performed to obtain principal components (PCs). The PCs could be expressed as

$$z_{ij} = \alpha_{i1}x1_j + \alpha_{i2}x2_j + \alpha_{i3}x3_j + \alpha_{i4}x4_j + \alpha_{i5}x5_j + \alpha_{i6}x6_j \quad (1)$$

where

z is the component score

α is the component loading
x_1 is sulphur dioxide (SO_2)
x_2 is particulate matter below 10 microns (PM_{10})
x_3 is ozone (O_3)
x_4 is nitrogen dioxide (NO_2)
x_5 is carbon monoxide (CO)
x_6 particulate matter below 2.5 microns ($PM_{2.5}$)
i is component number
j is the sample number

The correlation of variables in a PCA was evaluated using Bartlett's test of sphericity. The null hypothesis (H_0) suggested the original correlation matrix was an identity matrix, indicating unrelated variables. The alternative hypothesis (H_1) suggested the correlation matrix was significantly different from the identity matrix. A significant result (p-value < 0.05) indicated that factor analysis might be useful for the data set [8]. According to [9], this approach was used to describe the variability of a large group of connected variables and a smaller set of uncorrelated variables known as principal components. The KMO test was a measure designed to determine the suitability of data for factor analysis, measuring the adequacy of sample size. The test examined the sampling adequacy for each variable in the model and the whole model. KMO values below 0.6 indicated inadequate sampling and required corrective action [8].

Varimax rotation is a method used to reduce component complexity in PCs generated by the PCA, simplifying and improving the examination of variable structure. It involves raising high loads while decreasing small loads. PCs with eigenvalues greater than one were used to construct new variables, such as varimax factors (VFs) or factor loads, specifically in the varimax rotation process [9]. The Kaiser Criterion addresses maintain the required number of components. Varimax rotations require several variables aligned with conventional criteria, including non-observable, theoretical, and latent variables. Varimax factors (VFs) were the values used to calculate the correlation between variables. The values of varimax factors (VFs) could be categorised into three groups as shown in Table 2. The selection standard for this analysis is determined for VFs with absolute values over 0.75 [9].

Table 2: Category of varimax factor values

Varimax Factor Values	**Category**
> 0.75	Strong
$0.50 \geq VF \geq 0.75$	Moderate
$0.30 \geq VF \geq 0.49$	Weak

Result and Analysis

Data Processing

Multiple imputation was the method used to address the missing values. Since Klang data contained missing values, multiple imputations were performed.

Table 3: Result Variables

Result Variable	N of Valid Cases
Particulate matter below 10 microns (PM_{10})	1096
Particulate matter below 2.5 microns ($PM_{2.5}$)	1096
Sulphur dioxide (SO_2)	1096
Nitrogen dioxide (NO_2)	1096
Ozone (O_3)	1096
Carbon monoxide (CO)	1096

After processing the data using IBM SPSS Statistics, it was found that the data from Klang had missing values. In Klang, the variable nitrogen dioxide (NO_2) had two observations with missing values, while carbon monoxide (CO) had one observation. Since the data is numerical data, multiple imputations were conducted by replacing the missing values with the mean value of each variable. Table 3 presents the variables from the data in Klang after the missing values problem was resolved.

Dimensionality Reduction

After analyzing the data using IBM SPSS Statistics, the result revealed which pollutants were significant for determining the air pollution index for Klang. Table 4, presents the results of Kaiser-Meyer-Olkin (KMO) and Bartlett's sphericity tests for Klang.

Table 4: KMO and Bartlett's Test for Klang

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.652
Bartlett's Test of Sphericity	Approx. Chi-Square	3350.574
	df	15
	Sig.	0.000

The Kaiser-Meyer-Olkin (KMO) test for Klang data showed that the sampling adequacy was greater than 0.6, suggesting that all variables were suitable for further analysis. Factor loading calculations were conducted to evaluate the relationships between variables and extracted factors. Aside from that, Bartlett's test of sphericity displayed that the air pollution index data met the sphericity requirement, as the p-value was 0.000, which is less than 0.05. Consequently, the variables were determined to be connected rather than orthogonal. This indicated that PCA enabled the interpretation of data variability with less variables than the original number of variables.

Table 5 shows the results after applying varimax rotation. The number of principal components (PCs) to be retained was determined by examining the value of eigenvalues greater than one from the initial eigenvalues, which listed the six variables. According to [8], the retained components should account for at least 50% of the total variance, as recommended.

Table 5: Total Variance Explained for Klang

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variances	Cumulative %	Total	% of Variances	Cumulative %
1	**2.689**	44.825	44.825	2.686	44.765	44.765
2	**1.151**	19.191	64.016	1.155	19.250	64.016
3	0.987	16.456	80.471			
4	0.612	10.194	90.665			
5	0.511	8.524	99.189			
6	0.049	1.811	100.00			

As a result, Table 5 reveals that only two out of six principal components (PCs) were selected for Klang, as the eigenvalues were greater than one, accounting for 64.016% of the total variance in the data. Figure 1 shows the scree plots of the eigenvalues for the components in Klang, showing a decreasing trend as the number of components increases.

The study selected PC1 and PC2 for Klang due to their eigenvalues exceeding one, allowing for the creation of new variables known as varimax factors (VFs). VF values greater than 0.75 were chosen as a criterion for this study since it falls in the strong category, indicating moderate to significant highlights on eliminated components. Table 6 showed that three out of six air pollutants met the 0.75 VF threshold, with PM_{10}, $PM_{2.5}$ and O_3 being the most significant contributors.

Table 6: Rotated Component Matrix for Klang

	Component	
	1	2
PM_{10}	**0.933**	-0.127
$PM_{2.5}$	**0.920**	-0.136
CO	0.684	0.350
NO_2	0.632	0.456
O_3	0.299	**-0.826**
SO_2	0.112	0.330

After rotation for Klang, Component 1 accounted for approximately 44.765% of the variation in air quality data as shown in Table 5. Based on the varimax factor (VF) value categories outlined in methodology, Table 2, two pollutants with absolute VF values exceeding 0.75 have a strong VF value, which is PM_{10} (0.933) and $PM_{2.5}$ (0.920) for Klang. According to [10], particulate matter (PM), also known as dust or aerosols, is primarily caused by road traffic, construction, mining operations, wind erosion, agricultural activities, southwest monsoon wind, and burning biomass during the dry season in Malaysia.

Component 2, as shown in Table 5 contributes about 19.250% of the variability in air quality data. For this component, only one pollutant that contained absolute values of VF that exceed than 0.75: O_3 (-0.826) for Klang. According to [11], the rapid growth of industries and human activities in the ground-level atmosphere leads to increased levels of nitrogen oxides (NOx) and volatile organic compounds (VOC), precursors to ozone. Motor vehicles are a significant contributor to air pollution, public health concerns, and issues linked to the changing global climate.

Conclusion

In conclusion, the primary air contaminants in Klang may be determined using principal component analysis (PCA). The data on air pollution satisfied the sphericity assumption of Bartlett's test, with a p-value of less than 0.05, indicating that the variables associated with air pollution were correlated rather than orthogonal. By using the Kaiser-Meyer-Olkin (KMO) test, it was determined that all pollutants considered worthy of further investigation were acceptable and that the sampling adequacy was higher than 0.6. Two different PCs produced by rotating the PCA with varimax rotation showed that three out of six pollutants are the most significant pollutants in Klang, with the selected pollutants namely PM_{10}, $PM_{2.5}$ and O_3. The identification of these key pollutants is significant, as it provides a clear direction for policymakers, environmental agencies, and public health authorities to design targeted mitigation strategies. Such measures can reduce exposure to harmful pollutants, lower the prevalence of pollution-related diseases, and promote both public health and environmental sustainability in Selangor.

Acknowledgements

First and foremost, a big thank you to everyone who helped on this project. Your efforts, regardless of how small, have been important to the success of this study. A sincere thank you goes to the Universiti Teknologi MARA, Seremban Campus, Negeri Sembilan branch, for their important assistance and help which made the teamwork required for this project much simpler. The Department of the Environment (DOE) deserves credit on the tireless efforts to share accurate and comprehensive data, which significantly enhanced the accuracy of the study and results.

References

[1] World Health Organization. (2023). Monitoring air pollution levels is key to adopting and implementing global air quality guidelines. https://www.who.int/news/item/10-10-2023-monitoring-air-pollution-levels-is-key-to-adopting-and-implementing-who-s-global-air-quality-guidelines.%20Accessed%202024%20April%2027

[2] Ku Yusof, K., Azid, A., Sani, M. S. A., Samsudin, M. S., Amin, S., Rani, N., & Jamalani, M. A. (2019). The evaluation on artificial neural networks (ANN) and multiple linear regressions (*mlr*) models over particulate matter (*pm*10) variability during haze and non-haze episodes: A decade case study.

[3] Astro Awani (2024). Vehicle emissions are polluting malaysia's cities. https://www.astroawani.com/berita-malaysia/vehicle-emissions-are-polluting- malaysias-cities-453313

[4] Mohamed, N., Sulaiman, L. H., Zakaria, T. A., Kamarudin, A. S., & Rahim, D. A. (2016). Health risk assessment of *pm*10 exposure among school children and the proposed API level for closing the school during haze in Malaysia. *International Journal of Public Health Research*, *6*(1), 685–694.

[5] Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Tahmasebi Birgani, Y., & Rahmati, M. (2019). Air pollution prediction by using an artificial neural network model. *Clean Technologies and Environmental Policy*, *21*, 1341–1352.

[6] Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100.

[7] Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1), 20–30.

[8] Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, 9(1), 4–11.

[9] Mahmud, N., Zulkifli, N. E. S., & Muhammad Pazi, N. S. (2021). A study on air pollution index in Sabah and Sarawak using principal component analysis and artificial neural network. *Voice of Academia (VOA)*, 17(1), 20–29.

[10] Sentian, J., Jemain, M. A., Gabda, D., Franky, H., & Wui, J. C. H. (2018). Long-term trends and potential associated sources of particulate matter (PM_{10}) pollution in Malaysia. *WIT Transactions on Ecology and the Environment*, *230*, 607–618.

[11] Hashim, N. I. M., Yusoff, N. A. I. M., Noor, N. M., & Ul-Saufie, A. Z. (2019). Assessment of surface ozone concentration in northern peninsular malaysia. *IOP Conference Series: Materials Science and Engineering*, 551(1), 012100.